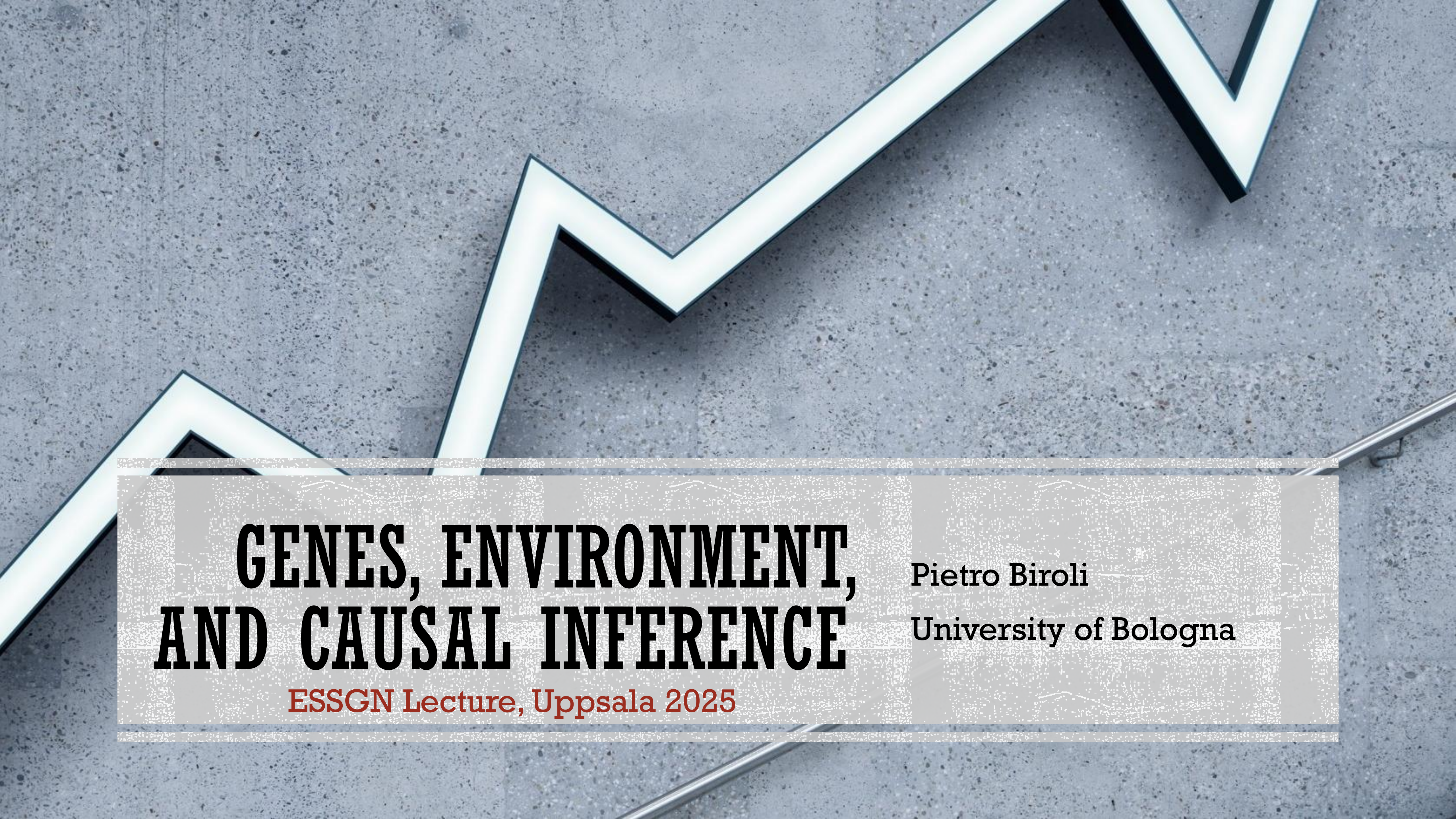


GENES, ENVIRONMENT, AND CAUSAL INFERENCE

ESSGN Lecture, Uppsala 2025

Pietro Biroli

University of Bologna

GAME PLAN

- Understand "Causal Inference"
  - Discuss common evaluation problems
  - Distinguish good from bad evaluation

- Review methods of evaluation
  1. RCT
  2. Diff-in-Diff
  3. RDD

- Estimate Gene-by-Environment Interplay (GxE)

INTRO

GAME PLAN

- Understand "Causal Inference"
  - Discuss common evaluation problems
  - Distinguish good from bad evaluation
- Review two methods of evaluation
  1. RCT
  2. Diff-in-Diff
  3. RDD
- Estimate Gene-by-Environment Interplay (GxE)

TWO TYPES OF "CAUSAL" EFFECTS

1. Effects of causes: *forward* causal inference
   - what happens if we do X?
   - e.g. what are the effects of smoking, schooling, advertisement

2. Causes of effects: *reverse* causal inference
   - what causes Y ?
   - why someone is in poor health, earns a lot, buys nutella?

DEFINITION OF CAUSALITY

I don’t have one. But two important components

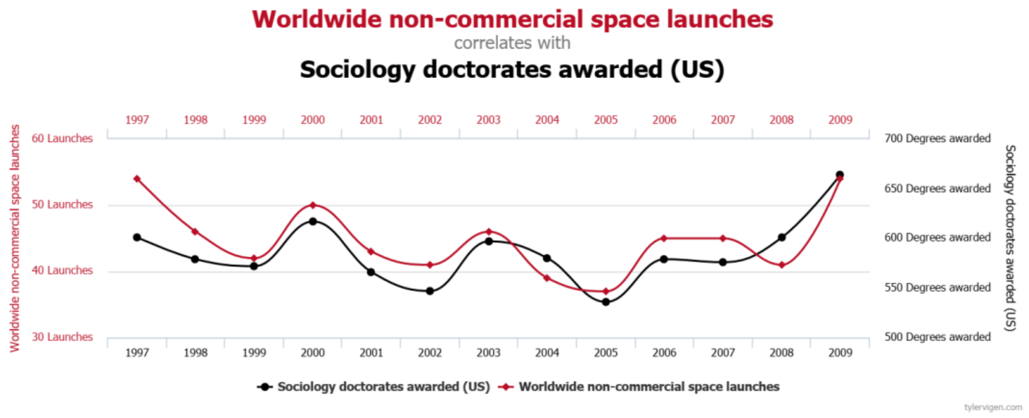
1. Theoretical **model** of counterfactuals
2. **Manipulation**

- Causality: property of a model following some rules
  - e.g. laws of physics, utility maximization, rules of social interaction
  - The more precise and articulated the model, the more precise the definition of causality
  - See [Mill, 1848, Marshall, 1890, Haavelmo, 1943, Holland, 1986, Heckman, 2005]
  - [Pearl, 2000, Pearl and Mackenzie, 2018] alternative approach

GWYNETH PALTROW
What if one split second sent your life...
SLIDING DOORS
...in two completely different directions?
JEDEN TAG, JEDE SEKUNDE TRIFFST DU EINE ENTSCHEIDUNG DIE DEIN LEBEN VERÄNDERN KANN
LOLA RENNT
mit FRANKA POTENTE und MORITZ BLEIBTREU
ein Film von TOM TYKWER
DVD
"GRIPPING TO THE END."
– David Denby, The New Yorker
uncertainty
joseph gordon-levitt
lynn collins
Life or death is a coin toss.

MODELING COUNTERFACTUAL

- If only I could go back and do it again …



COMMON ERRORS TO AVOID

Correlation is NOT causation

Confuse variation in outcomes with "impact"

3 TASKS OF CAUSAL ANALYSIS

1. **Define** the set of **counterfactuals**
   - Requires: a scientific theory
   - It's a matter of logic, convention, and imagination
2. **Identify** parameters from population data
   - Requires: mathematics of point or set identification
   - Find a unique mapping from population moments to the parameters
3. **Estimate** parameters from real data
   - Requires: estimation and testing theory
   - Statistical inference, considering sampling variation and data limitations

POTENTIAL OUTCOMES MODEL

- $Y_{i,1}$: potential outcome when treated $E(Y_i|D_i = 1)$
- $Y_{i,0}$: potential outcome when not treated $E(Y_i|D_i = 0)$

- Only one outcome is observed: $Y_i = D_i Y_{i,0} + (1 - D_i) Y_{i,1}$
- Holy grail: $\Delta_i = Y_{i,1} - Y_{i,0}$

WHAT WOULD WE WANT TO ESTIMATE?

- The proportion of people taking the program who benefit from it:
  - $\Pr(Y1 > Y0 | D = 1) = \Pr(\Delta > 0 | D = 1)$
- The proportion of the total population benefiting from the program:
  - $\Pr(Y1 > Y0 | D = 1)\Pr(D = 1) = \Pr(\Delta > 0 | D = 1)\Pr(D = 1)$
- The distribution of gains at selected base state values:
  - $F(\Delta | D = 1, Y0 = y0)$
- The voting criterion, i.e the share with ex-ante net benefit:
  - $\Pr\ (IE\ (Y1 - Y0 - C > 0 | I))$
- The increase in the proportion of outcomes above a certain threshold y due to a policy:
  - $\Pr(Y1 > y | D = 1) - \Pr(Y0 > y | D = 1)$

WHAT DO WE USUALLY ESTIMATE?

- (Conditional) Average Treatment Effect: (C)ATE

$$\mathrm{E}(\Delta_i|X) = E(Y_{i,1} - Y_{i,0}|X)$$
$$= E(Y_{i,1}\ |X) - E(Y_{i,0}|X)$$

- Why?
  - Additive separability,
  - Easier to estimate, under some assumptions (e.g RCT)

LIMITATIONS

- Binary treatment
  - Potential extensions to multivalued/continuous treatments [Lee and Salani, 2015]
- **SUTVA:** stable unit treatment value assumption
  - Potential outcomes depend on own treatment only
  - No spill-overs across i, no general equilibrium effects
    - Note: *i* could be a group!

PROBLEMS

GAME PLAN

- Understand "Causal Inference"
  - Discuss common evaluation problems
  - Distinguish good from bad evaluation
- Review two methods of evaluation
  1. RCT
  2. Diff-in-Diff
  3. RDD
- Estimate Gene-by-Environment Interplay (GxE)

MAIN PROBLEMS OF PROGRAM EVAL

Two main statistical problems related to the causal evaluation of a program:

1. The missing counterfactual
2. The selection problem

1. THE MISSING COUNTERFACTUAL

We observe only one of the two potential outcomes for each individual

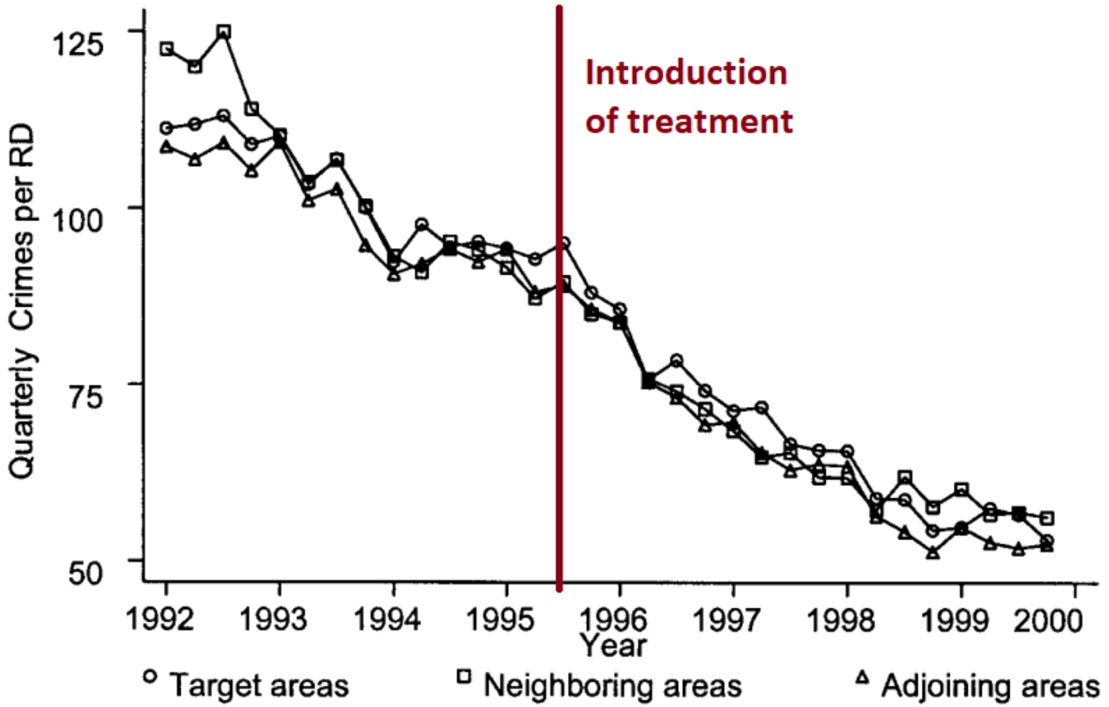
- The unobserved outcome is called the "Missing Counterfactual"
- Impossible to determine the impact of treatment without this counterfactual
- Causal *inference*: how *estimate* the missing counterfactual

COMMON MISTAKE: BEFORE/AFTER COMPARISON

- Measure outcome before and after treatment
- Use pre-treatment as a proxy for missing counterfactual
- Calculate difference: (post)-(pre)
- Call it "effect of treatment"

WRONG!

  - What would have happened over time if treatment weren't there??



BEFORE/AFTER COMPARISON

- Ignores the natural evolution of the outcome
  - Sometimes disputes between partners exacerbate or resolve, regardless of access to a lawyer
- Only plausible if outcome is constant over time
  - Maybe height?
- Very very UNLIKELY to happen in your case

2. THE SELECTION PROBLEM

Who are the people in the treatment group?

  - Very selected, non-representative sample!

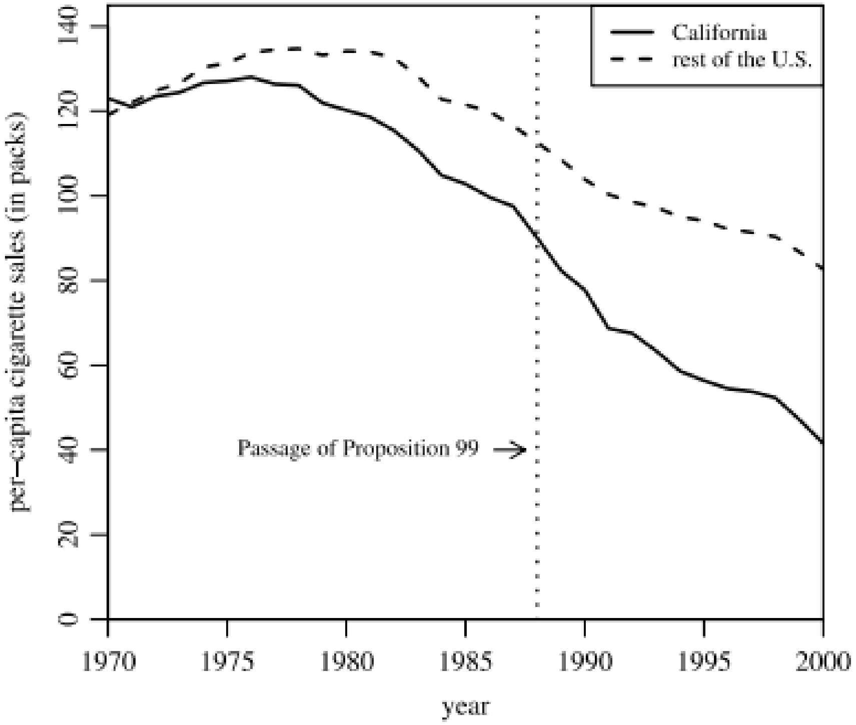
- Participants choose to become part of treatment
  - Based on obs and unobs characteristics, which also drive outcomes
- Participants are different from:
  - all of the eligible population
  - those who did not get treated
  - themselves prior to the start of the program

COMMON MISTAKE: TREAT/CONTROL COMPARISON

- Measure outcome for people who are never treated (control)
- Use control-group outcomes as proxy for missing counterfactual
- Calculate difference: (treat)-(control)
- Call it "effect of treatment"

WRONG!

- What would have happened to people who decided to be treated if treatment weren't there??



TREAT/CONTROL COMPARISON

Why did certain people get treated while others did not?

- Because of the selection problem, the underlying assumptions of the treatment/control comparison are questionable.
- The outcomes for untreated individuals are likely to be a bad estimate for the counterfactual outcomes of the treated individuals.
  - E.g. people who were treated by a doctor and those who weren’t

TWO TYPES OF SELECTION

- Selection on Observables:
  - participants are different from non-participants in something that we can observe and measure:
    - Plaintiffs' age, gender, income
- Selection on Unobservables:
  - participants are different from non-participants in terms of something we **cannot** observe or measure:
    - Recklessness, morality, motivation, trustworthiness

SOLUTIONS

WHAT TO DO?

- Selection on observables can be accounted for by using statistics
  - E.g. control for those variables in a regression or matching algorithm
- Selection on unobservables: cannot be easily solved
- ALWAYS think and ask: how are treated and control people different? Can we measure all of these?
- If NOT: look for an identification strategy

IDENTIFICATION STRATEGY

- Clever design and use of data to estimate the missing counterfactual and overcome selection problem
- Common ID strategies:
  1. Randomized Controlled Trials (RCT)
  2. Difference in differences (Diff-in-diff)
  3. Regression Discontinuity Design (RDD)
  4. Instrumental Variable Regression (IV)

GAME PLAN

- Understand "Causal Inference"
  - Discuss common evaluation problems
  - Distinguish good from bad evaluation
- Review two methods of evaluation
  1. RCT
  2. Diff-in-Diff
  3. RDD
- Estimate Gene-by-Environment Interplay (GxE)

RANDOMIZED CONTROLLED TRIALS

The Experimental Revolution:

- Clinical Trials in Medicine
  - 1774, James Lind and lemons to cure of scurvy
  - COVID-19 vaccines
- Economists jumped on the ship
  - Law and economics
  - Education econ
  - Urban econ
  - Development econ: 2019 Nobel Prize + Duflo TED talk

THE RATIONALE BEHIND RCT

If treatment status is randomly determined:

- Observable *and unobservable* characteristics are balanced (~ equal) for treated and control group

➡ No selection problem

➡ Control group ~ missing counterfactual

$E(X,U|D=1) \sim E(X,U|D=0)$

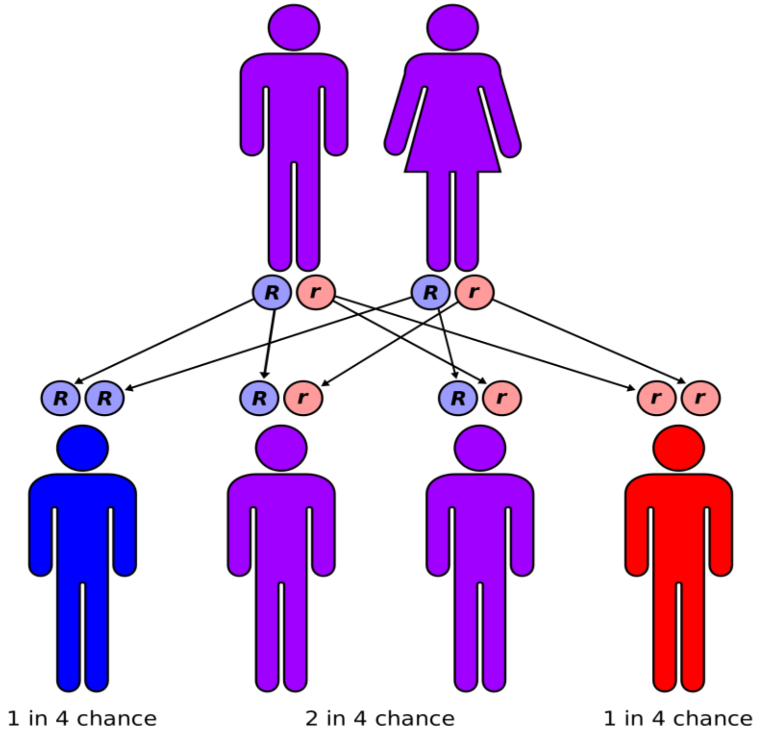
ESTIMATION WITH RCT

1. Average outcome for treated group
2. Average outcome for control group
3. Take the difference

→ Estimate of the (Conditional) Average Treatment Effect

GENES AS "NATURAL" RCT?

- **"Mendelian Randomization"**
  - Each genotype has two alleles (two copies of each chromosome)
  - Inherit one chromosome from dad and one from mom
  - Which one you inherit happens *at random*
  - Must condition on parental genotype



GENETIC COUNTERFACTUAL

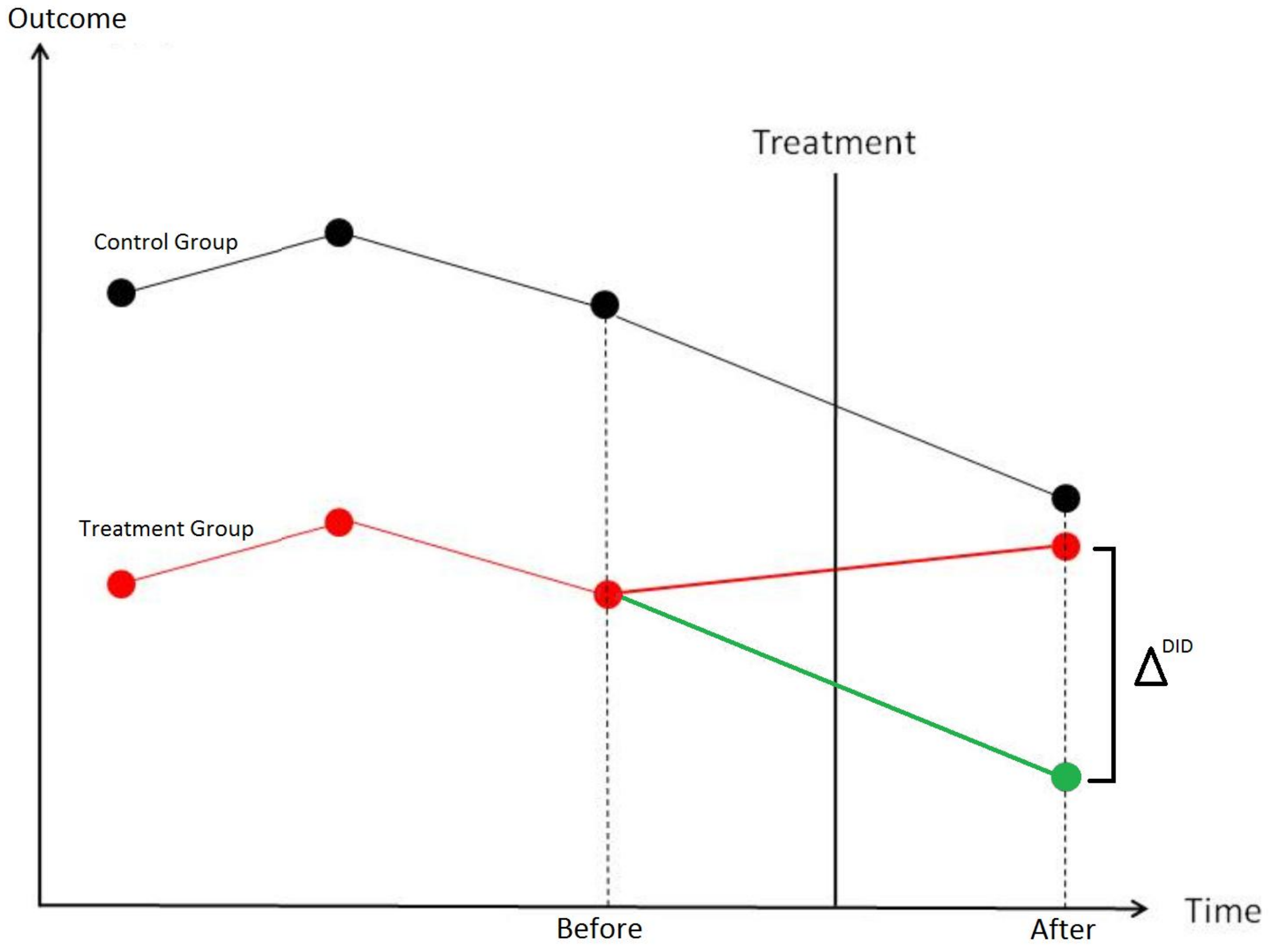
- Q: what is the “counterfactual” in genetics? Y0 vs Y??
  - SNP-level
  - PGI-level
- Q: what about intergenerational?
- Q: is this just a thought experiment?
  - CRISPR-Cas
  - Embrio selection

GAME PLAN

- Understand "Causal Inference"
  - Discuss common evaluation problems
  - Distinguish good from bad evaluation
- Review two methods of evaluation
  1. RCT
  2. Diff-in-Diff
  3. RDD
- Estimate Gene-by-Environment Interplay (GxE)

DIFFERENCE IN DIFFERENCES

- No randomization
  - Feasibility, ethical, time constraints
- However: there is data on
  - Treatment group: before and after intervention
  - Control group: on the same time period
- Can leverage diff-in-diff design:
  - Compare evolution of outcomes between T and C



DIFF-IN-DIFF

1. Parallel trends:
   - Check that T and C groups have similar trends *before* the treatment happens
2. Project Forward
   - the evolution of C group onto the T group
3. Difference from projected trend = estimated average treatment effect

MAIN ASSUMPTION: evolution of outcomes between T and C would have been the same in the absence of treatment

DIFF-IN-DIFF

- Use evolution over time and across groups to overcome missing counterfactual and selection
- Stronger statistical assumptions
- A bit harder to estimate
- Allows for evaluation ex-post

GAME PLAN

- Understand "Causal Inference"
  - Discuss common evaluation problems
  - Distinguish good from bad evaluation
- Review two methods of evaluation
  1. RCT
  2. Diff-in-Diff
  3. RDD
- Estimate Gene-by-Environment Interplay (GxE)

REGRESSION DISCONTINUITY DESIGN

Special causal method that sometimes happens in the wild:

- Design requirement:
  - **Cutt-off rule**: people above an arbitrary cutoff are more likely to be treated
- Data requirement:
  - a **LOT** of data near the threshold

THE ASSIGNMENT VARIABLE

- Treatment depends on **one continuous** variable X
  - assignment, running, or forcing variable X
- Characteristics of *X*:
  - Continuous
  - Affects *discontinuously* the probability of treatment $Pr(D = 1 | X)$ at cutoff point X=*c*
  - Related to potential outcomes in a *continuous* way
    - $Y\_d = g(X)$, with $g(.)$ continuous at $X = c$
- RDDs take two forms: sharp and fuzzy.

SHARP RDD

- Treatment probability jumps form 0 to 1 at the cutoff
  - Age for driving/voting
  - Election

G.W. Imbens, T. Lemieux / Journal of Econometrics 142 (2008) 615–635

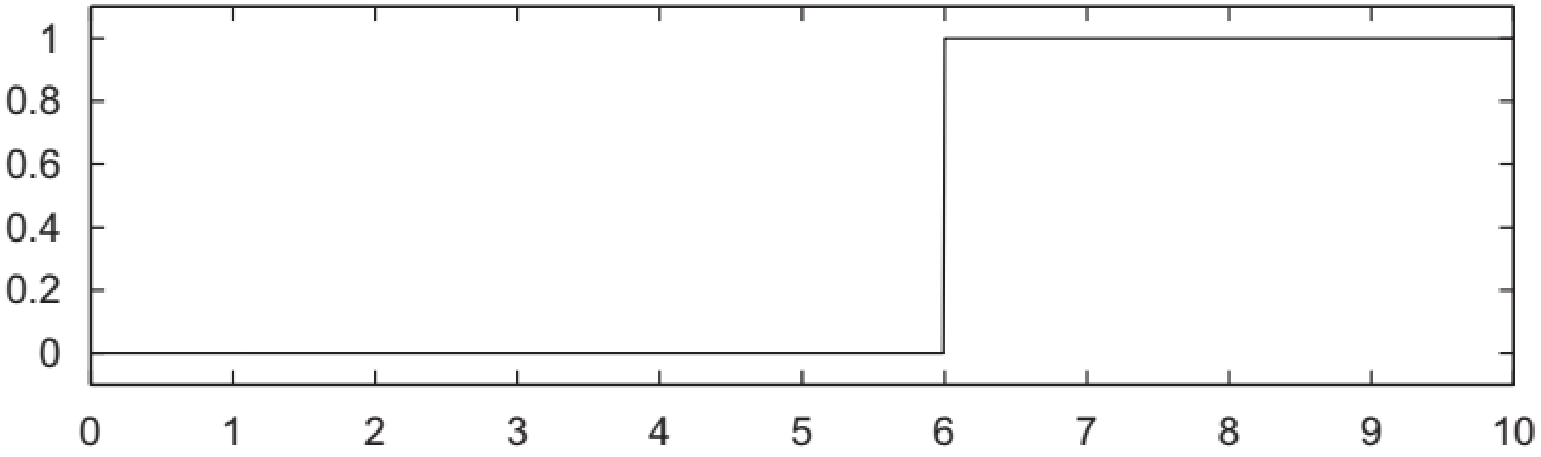


Fig. 1. Assignment probabilities (SRD).

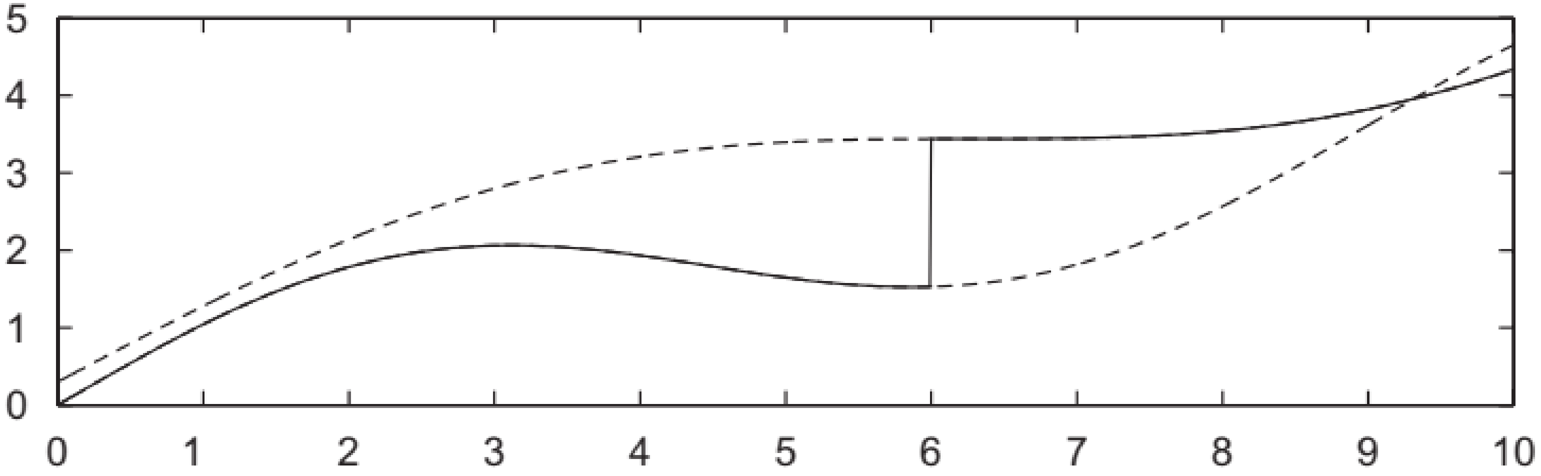


Fig. 2. Potential and observed outcome regression functions.

FUZZY RDD

- Treatment probability jumps by 0<k<1 at the cutoff
  - Driking age?

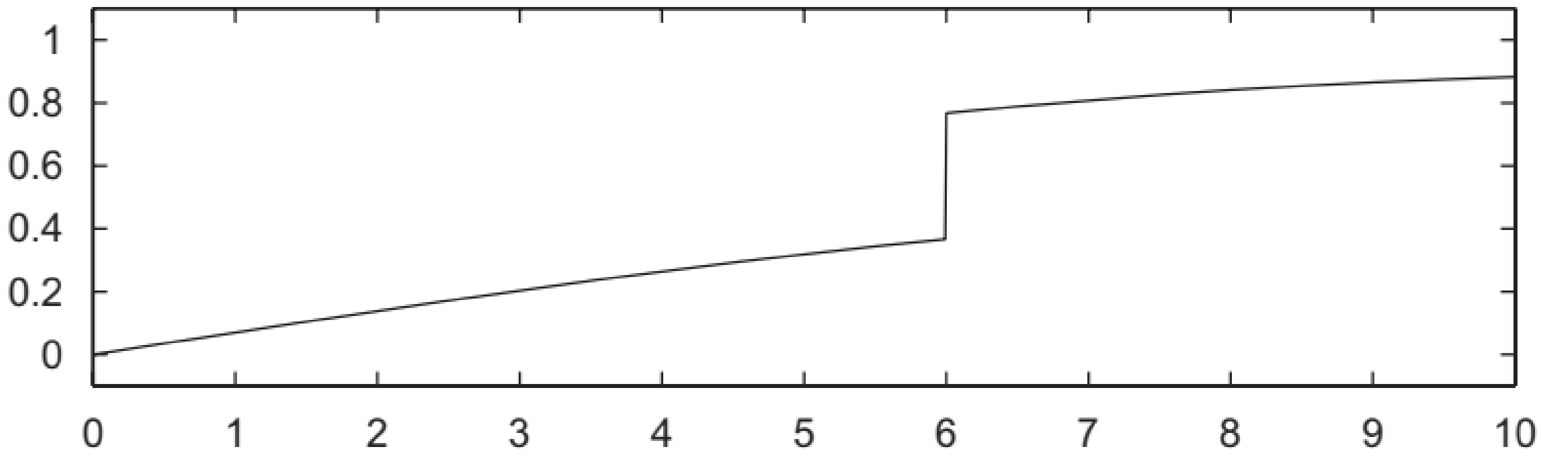


Fig. 3. Assignment probabilities (FRD).

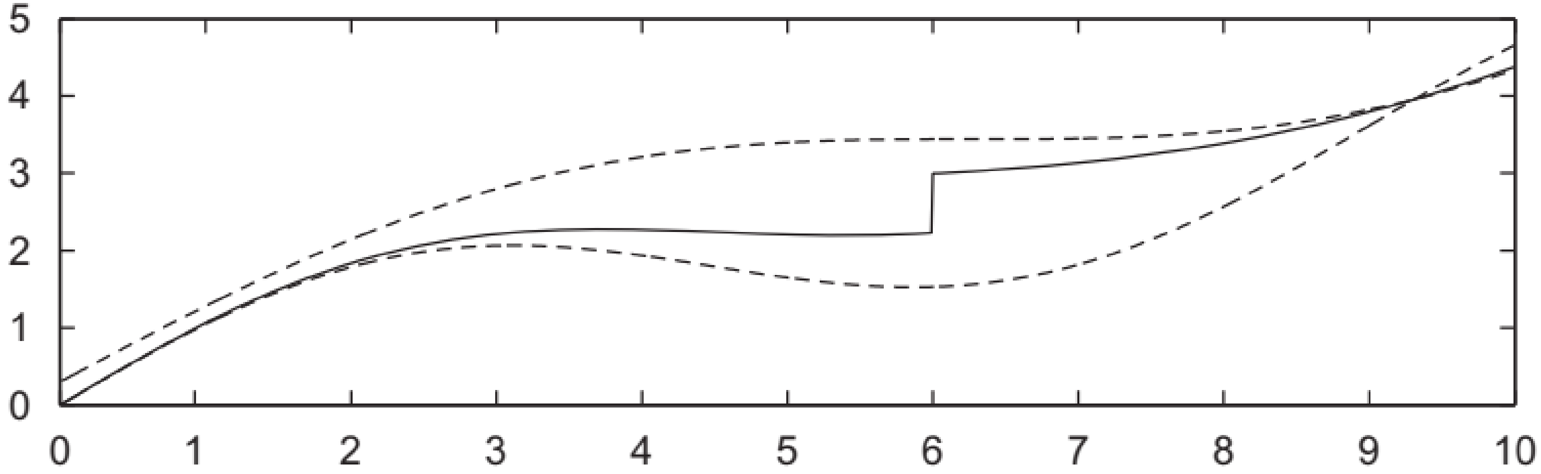
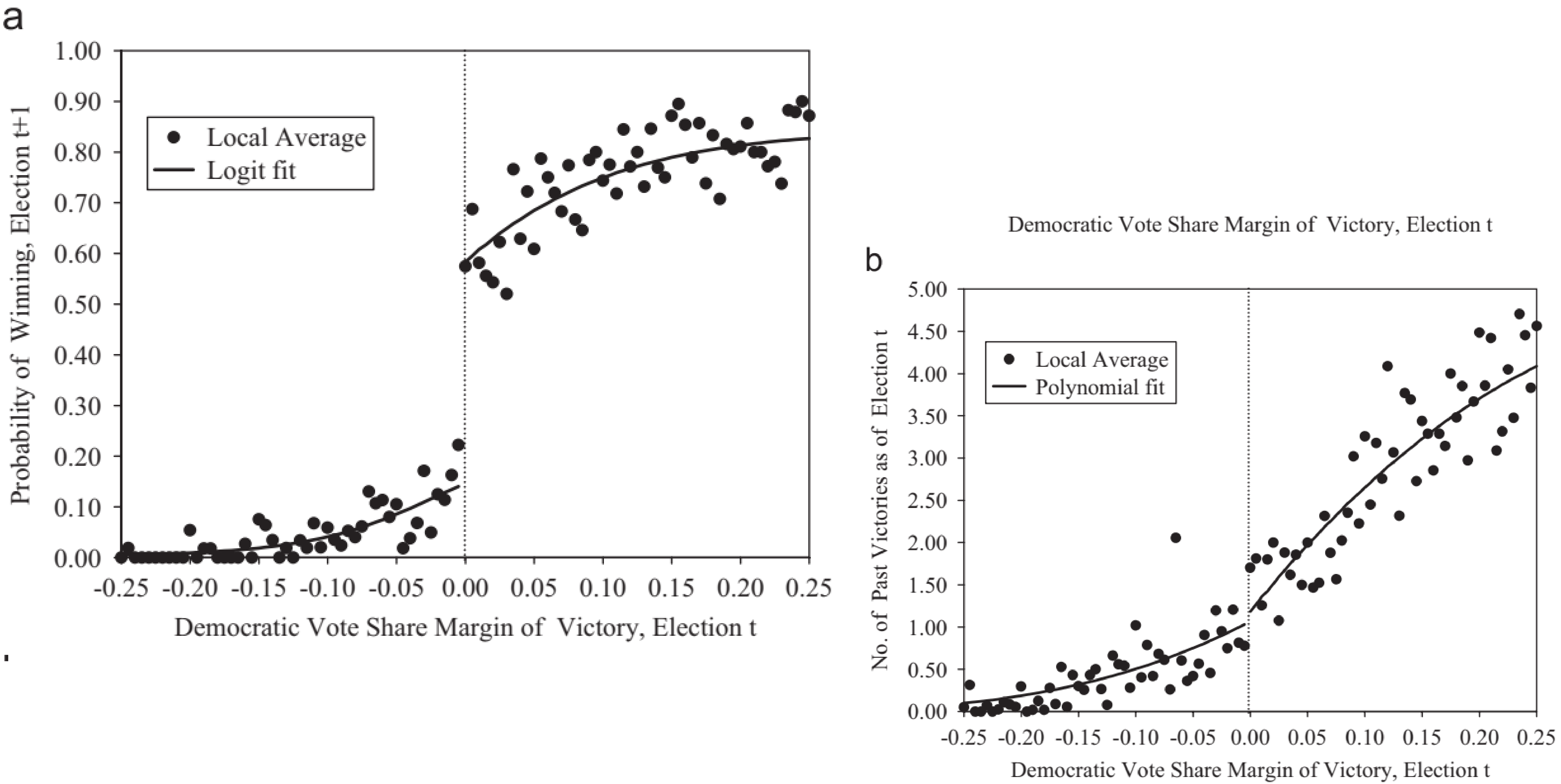


Fig. 4. Potential and observed outcome regression (FRD).

ESTIMATION

- Average difference in outcomes just before and after the cutoff
- Approximate this with a non-linear OLS (with a caliper)

$$Y = \beta_0 + \sum_k \beta_{low,k}(X_1 - c)^k + \delta D + \sum_k \beta_{upp,k}(X_1 - c)^k D + \varepsilon$$


a
Local Average
Logit fit
Probability of Winning, Election t+1
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-0.25
-0.20
-0.15
-0.10
-0.05
0.00
0.05
0.10
0.15
0.20
0.25
Democratic Vote Share Margin of Victory, Election t
Democratic Vote Share Margin of Victory, Election t
b
Local Average
Polynomial fit
No. of Past Victories as of Election t
5.00
4.50
4.00
3.50
3.00
2.50
2.00
1.50
1.00
0.50
0.00
-0.25
-0.20
-0.15
-0.10
-0.05
0.00
0.05
0.10
0.15
0.20
0.25
Democratic Vote Share Margin of Victory, Election t

BREAK: PLAY GENETIC PINBALL

- https://purplesloth.itch.io/genetic-pinball

MAX PLANCK INSTITUTE
FOR HUMAN DEVELOPMENT

PURPLE SLOTH STUDIO

SCAN ME

GAME PLAN

- Understand "Causal Inference"
  - Discuss common evaluation problems
  - Distinguish good from bad evaluation
- Review two methods of evaluation
  1. RCT
  2. Diff-in-Diff
  3. RDD
- Estimate Gene-by-Environment Interplay (GxE)

ARE WE THE WAY WE ARE BECAUSE …

… we are born this way

nature / genes

… or we became this way?

nurture / environment

NATURE *VIA* NURTURE

MATT RIDLEY (2003)

Distinction of nature vs nurture is obsolete!

- Gene and environment *interplay*:
  - Same gene with different effects depending on environment (GxE)
  - Genes influence the environment we select into (rGE)
    - Elliott Joslin (1950s)

      "Genes load the gun. Lifestyle pulls the trigger."
    - Erik Turkheimer (2000)

      "The nature-nurture debate is over.

      The bottom line is that everything is heritable."
    - Rutter (2006)

GXE: NON-LINEARITY OF G AND E

- DGP: $Y_i = F(a^*, G_i, E_i, e_i)$
  - Y = outcome
  - $a^* = a^*(G_i, E_i, e_i)$ individual choices
  - e_i randomness
  - GxE = non-zero cross-partial of G and E

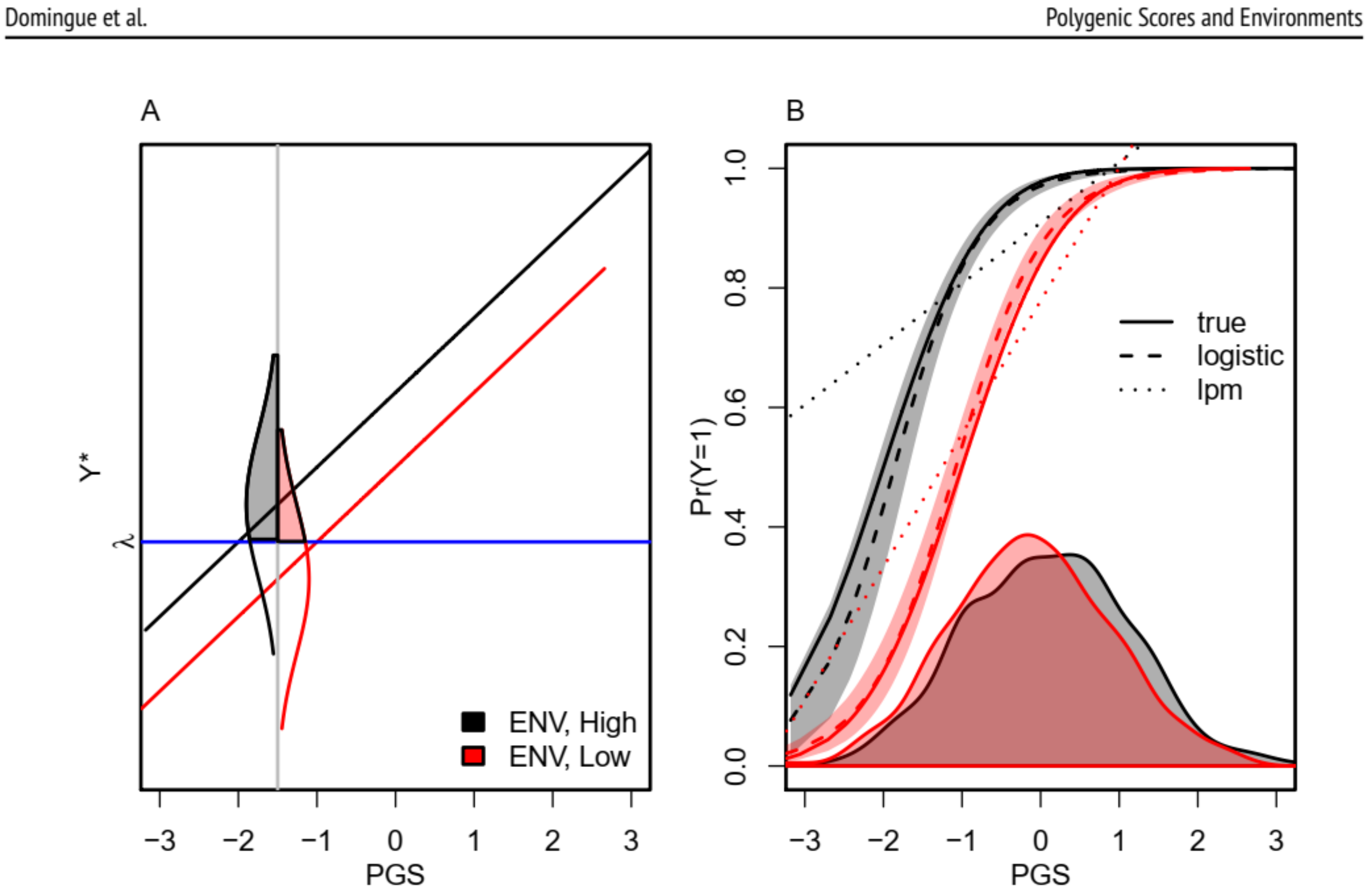
- Simplistic estimation:

$$Y_i = \alpha + \beta_G G_i + \beta_E E_i + \beta_{G\times E}\left(G_i \times E_i\right) + \theta E_i^2 + \rho G_i^2$$
$$+ \mu_x X_i + \mu_g\left(G_i \times X_i\right) + \mu_e\left(E_i \times X_i\right) + \varepsilon_i.$$

  - X_i predetermined controls, demeaned and interacted with G and E

SIDENOTE: BINARY OUTCOME

- Probit, logit, poisson regression etc: non-linear estimation
  - Very hard to estimate interactions properly



Domingue et al.

Polygenic Scores and Environments

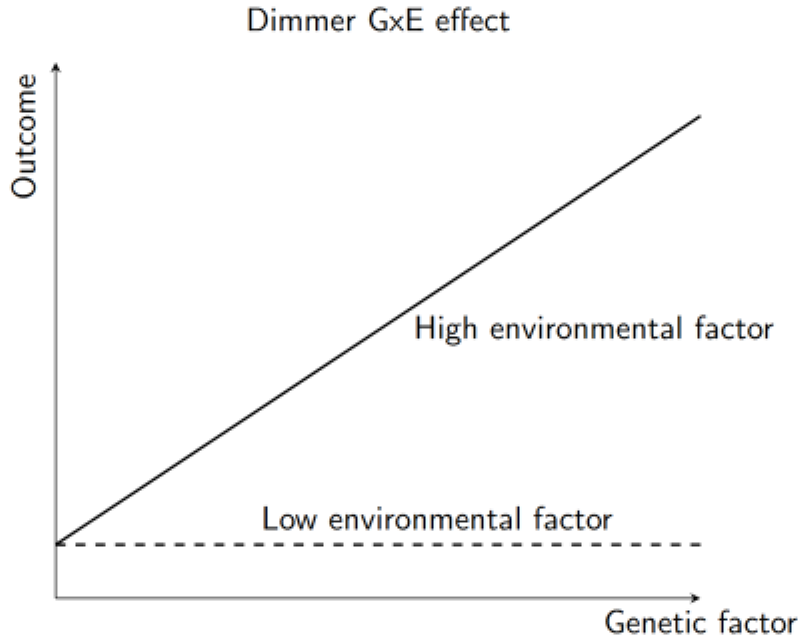
HOW TO MEASURE G?

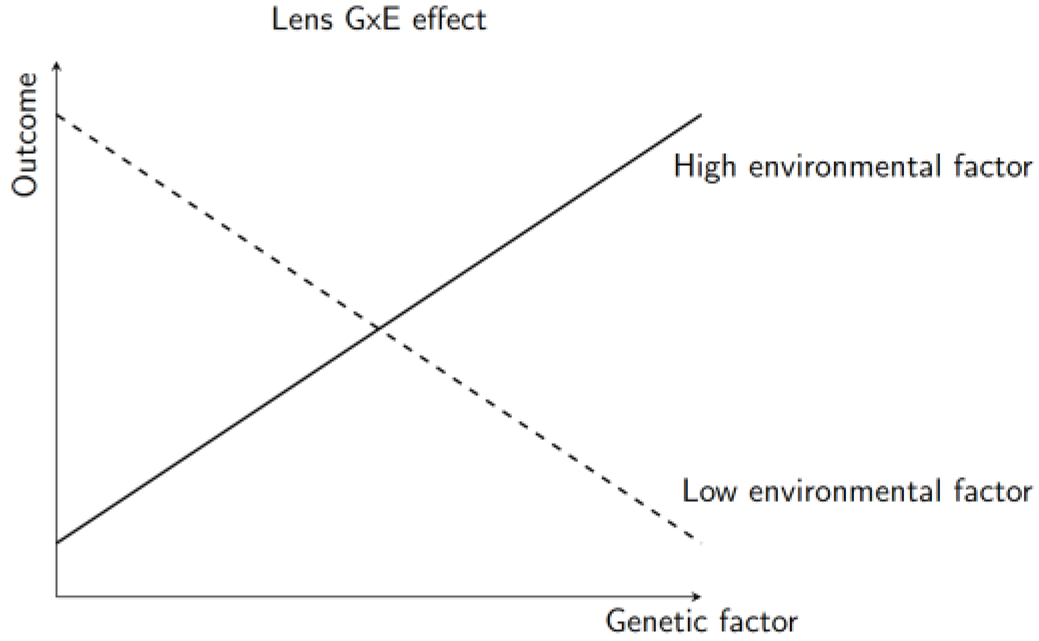
SNPs or PGI?

- SNP:
  - low predictive power
- PGI:
  - Measurement error
  - Re-introduce GWAS estimation problem: additivity; sample selection; wrong DGP
  - Which PGI phenotype? Mean, variance, bio-annotated?
- Check out Miao, Wu, Lu (2023) for more through discussion

TYPES OF GXE

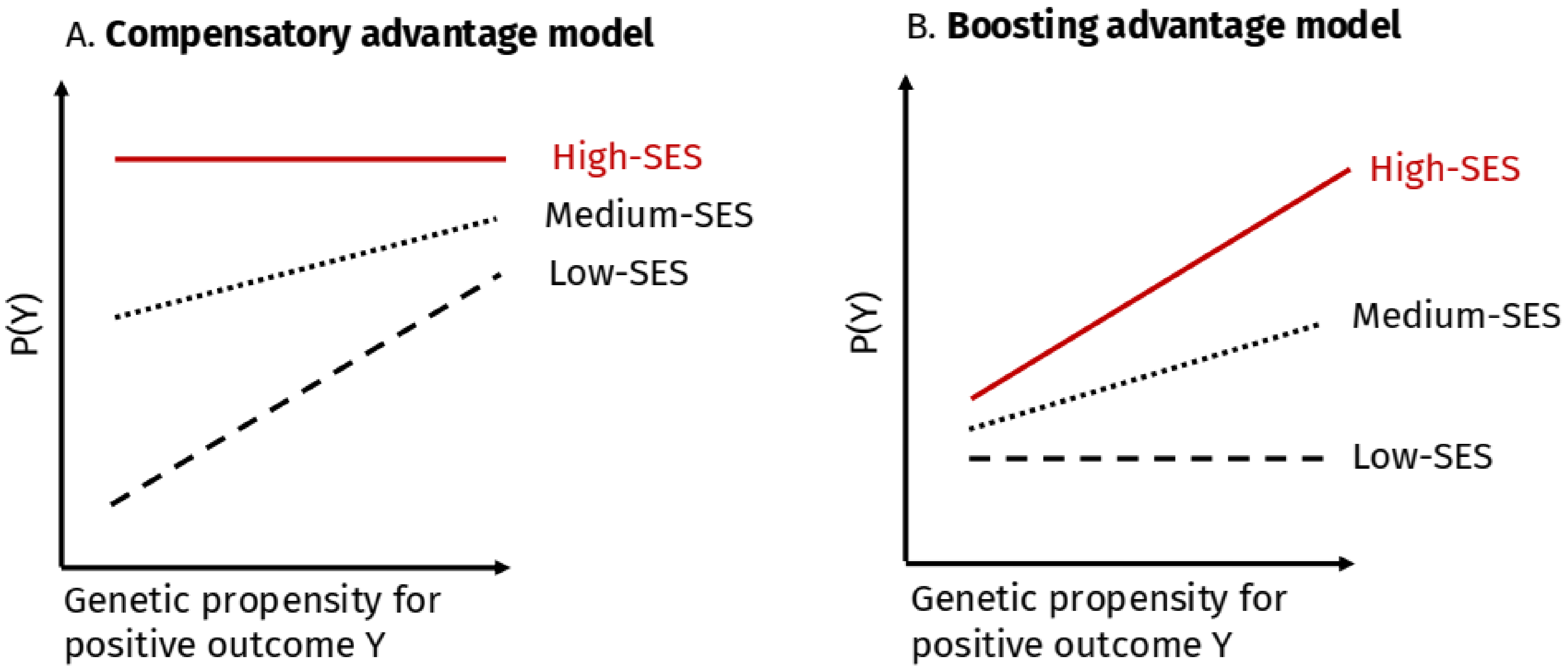
- Dimmer:
  - PGI has same sign, different slope
  - Rank-order preserving
- Lens:
  - PGI has different sign
  - Rank re-ordering
  - Mean ENV effect could be zero!





DIMMER: INCREASE OR REDUCE INEQUALITY?

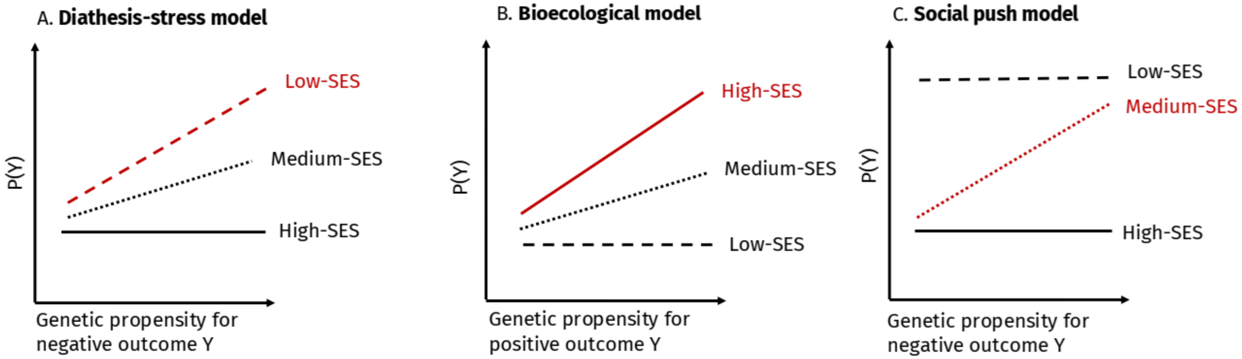
Figure 3: Main theoretical models to study GxSES in social stratification



Source: Gaia Ghirardi PhD Thesis, 2005

MORE CATEGORIES: GAIA GHIRARDI (2005)

Figure 2: Main theoretical models to study GxSES in behavioral genetics



Source: Gaia Ghirardi PhD Thesis (2005)

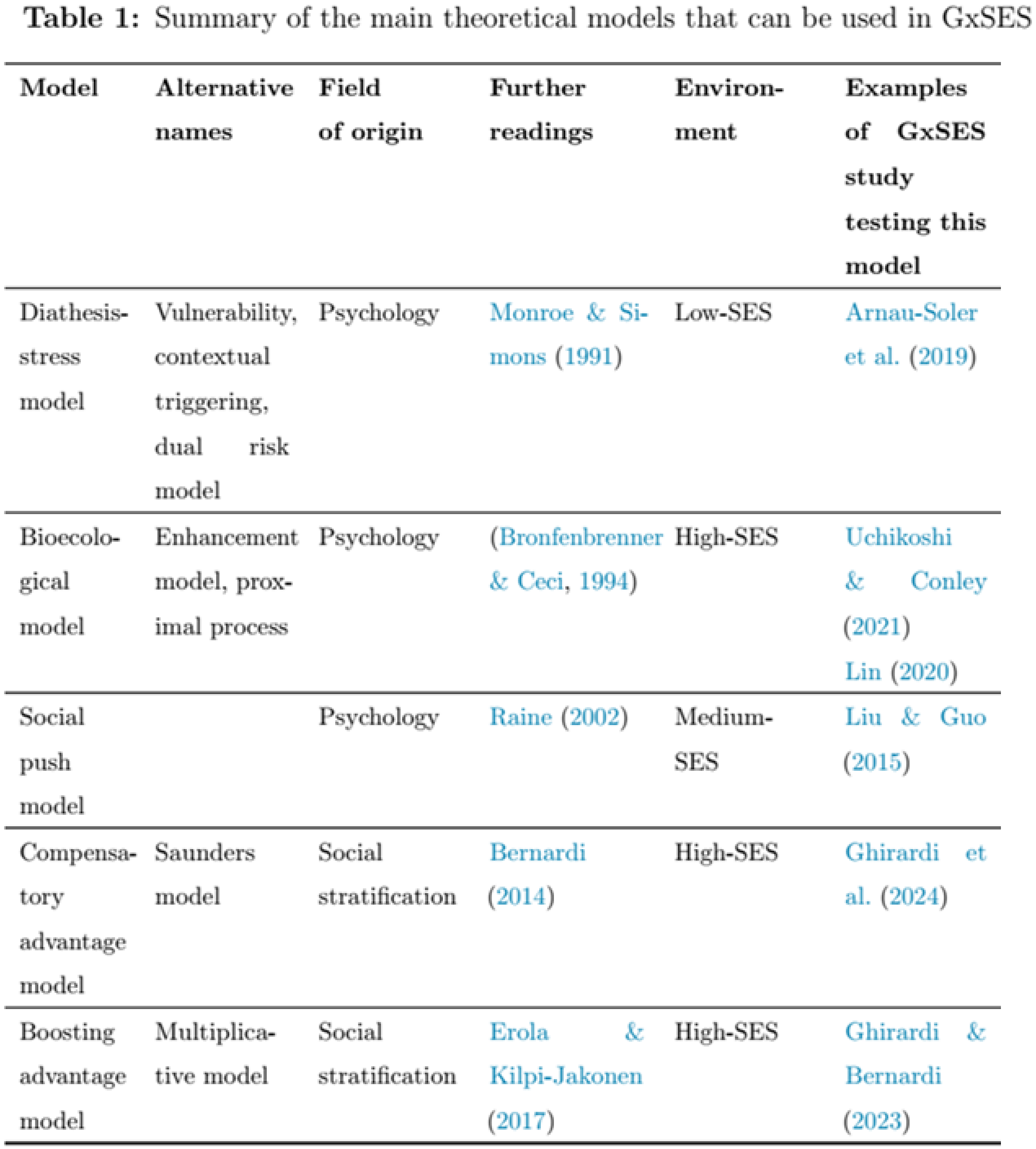
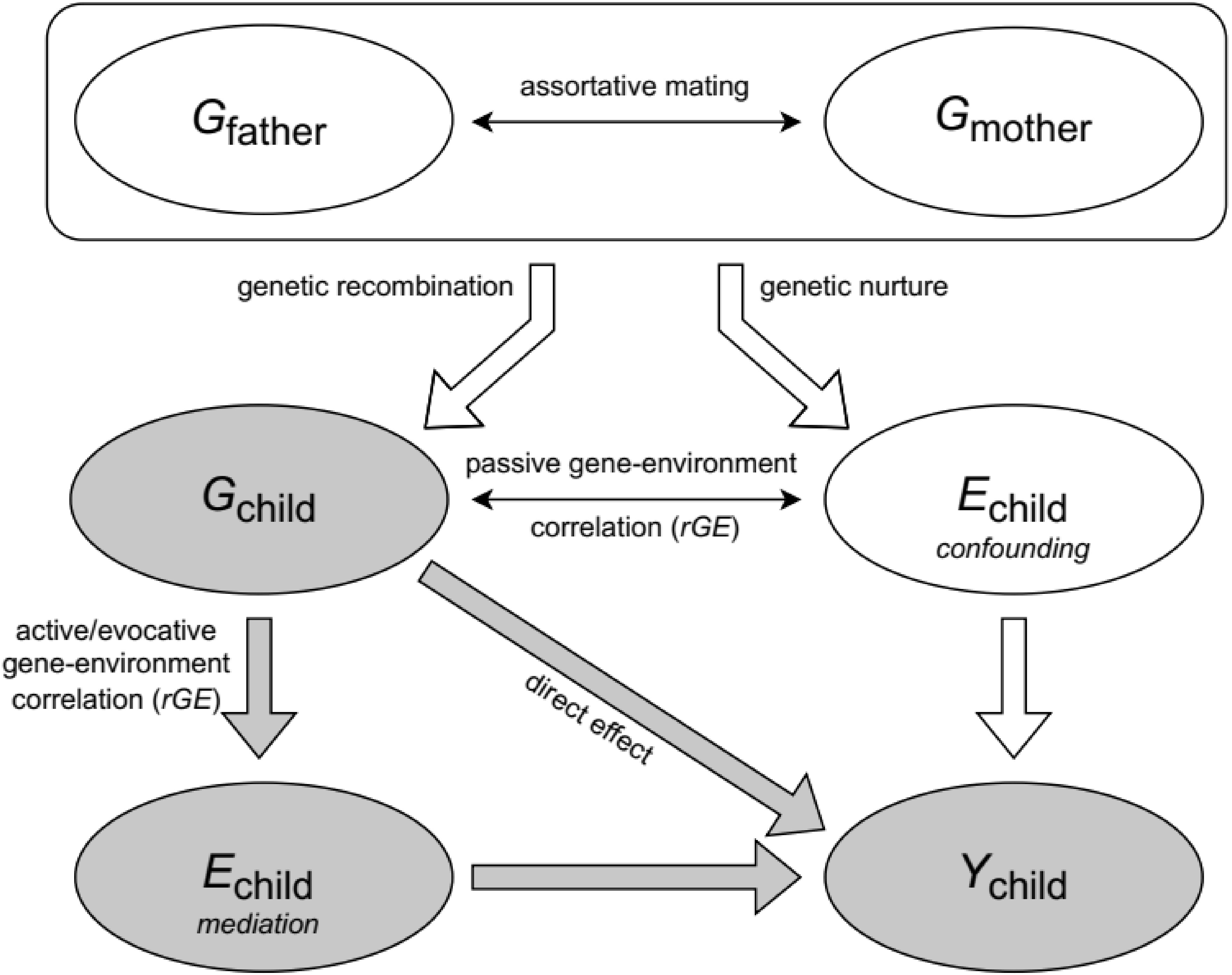


Table 1: Summary of the main theoretical models that can be used in GxSES

Model	Alternative names	Field of origin	Further readings	Environment	Examples of GxSES study testing this model
Diathesis-stress model	Vulnerability, contextual triggering, dual risk model	Psychology	Monroe & Simons (1991)	Low-SES	Arnau-Soler et al. (2019)
Bioecological model	Enhancement model, proximal process	Psychology	(Bronfenbrenner & Ceci, 1994)	High-SES	Uchikoshi & Conley (2021) Lin (2020)
Social push model		Psychology	Raine (2002)	Medium-SES	Liu & Guo (2015)
Compensatory advantage model	Saunders model	Social stratification	Bernardi (2014)	High-SES	Ghirardi et al. (2024)
Boosting advantage model	Multiplicative model	Social stratification	Erola & Kilpi-Jakonen (2017)	High-SES	Ghirardi & Bernardi (2023)

CAUSALITY AND CONFOUNDING

- Random inheritance from parents (Mendel)
- But genes do not live in a vacuum (Harden)



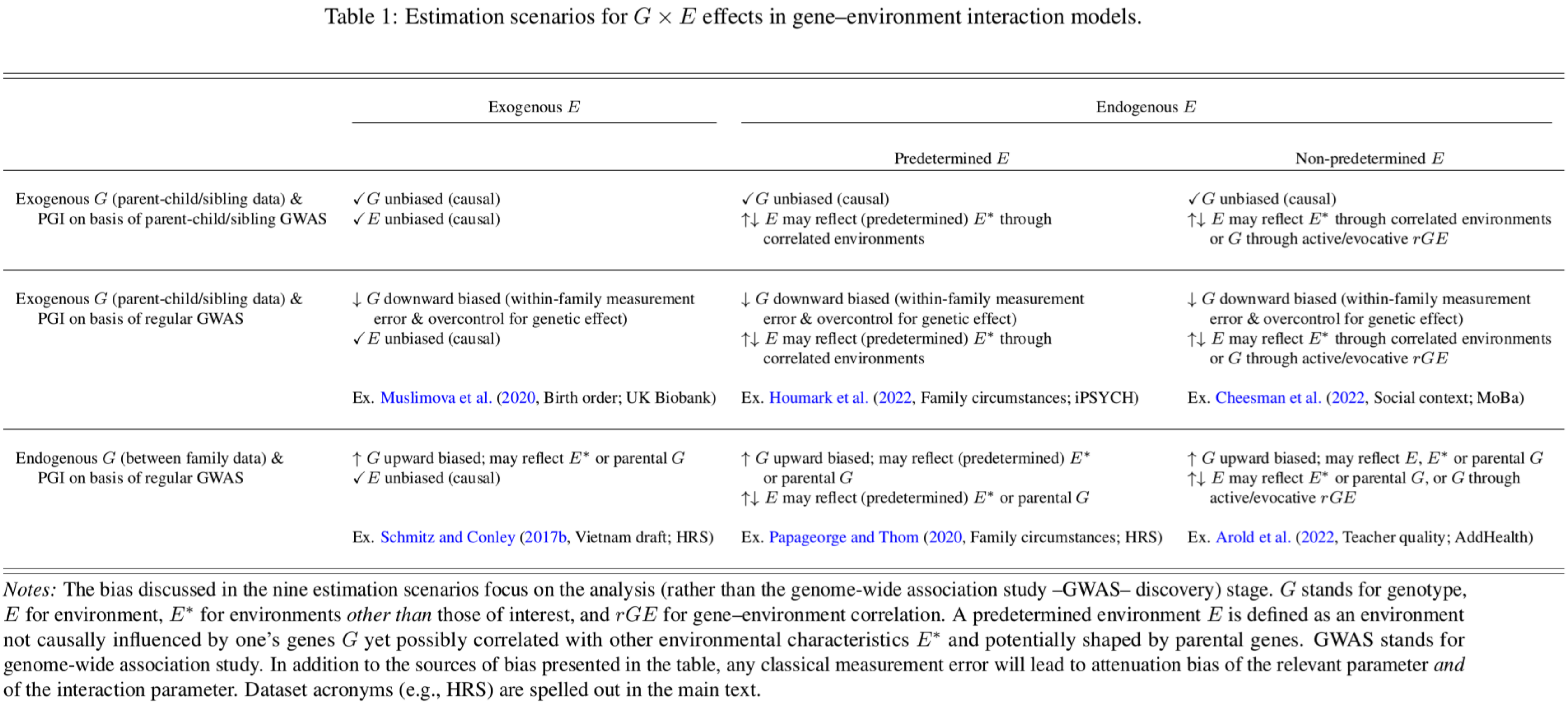


Table 1: Estimation scenarios for $G \times E$ effects in gene–environment interaction models.

	Exogenous E	Endogenous E	
		Predetermined E	Non-predetermined E
Exogenous G (parent-child/sibling data) & PGI on basis of parent-child/sibling GWAS	✓ G unbiased (causal) ✓ E unbiased (causal)	✓ G unbiased (causal) ↑↓ E may reflect (predetermined) E^* through correlated environments	✓ G unbiased (causal) ↑↓ E may reflect E^* through correlated environments or G through active/evocative rGE
Exogenous G (parent-child/sibling data) & PGI on basis of regular GWAS	↓ G downward biased (within-family measurement error & overcontrol for genetic effect) ✓ E unbiased (causal)	↓ G downward biased (within-family measurement error & overcontrol for genetic effect) ↑↓ E may reflect (predetermined) E^* through correlated environments	↓ G downward biased (within-family measurement error & overcontrol for genetic effect) ↑↓ E may reflect E^* through correlated environments or G through active/evocative rGE
	Ex. Muslimova et al. (2020, Birth order; UK Biobank)	Ex. Houmark et al. (2022, Family circumstances; iPSYCH)	Ex. Cheesman et al. (2022, Social context; MoBa)
Endogenous G (between family data) & PGI on basis of regular GWAS	↑ G upward biased; may reflect E^* or parental G ✓ E unbiased (causal)	↑ G upward biased; may reflect (predetermined) E^* or parental G ↑↓ E may reflect (predetermined) E^* or parental G	↑ G upward biased; may reflect E, E^* or parental G ↑↓ E may reflect E^* or parental G, or G through active/evocative rGE
	Ex. Schmitz and Conley (2017b, Vietnam draft; HRS)	Ex. Papageorge and Thom (2020, Family circumstances; HRS)	Ex. Arold et al. (2022, Teacher quality; AddHealth)

Notes: The bias discussed in the nine estimation scenarios focus on the analysis (rather than the genome-wide association study –GWAS– discovery) stage. G stands for genotype, E for environment, E^* for environments *other than* those of interest, and rGE for gene–environment correlation. A predetermined environment E is defined as an environment not causally influenced by one's genes G yet possibly correlated with other environmental characteristics E^* and potentially shaped by parental genes. GWAS stands for genome-wide association study. In addition to the sources of bias presented in the table, any classical measurement error will lead to attenuation bias of the relevant parameter *and* of the interaction parameter. Dataset acronyms (e.g., HRS) are spelled out in the main text.

GXE CHECKLIST

1. Power calculations
   - simulations
2. Check for rGE
   - Test of exogeneity
3. Functional form
   - Non-linear plots
4. GxE regression
   - Demeaned and interacted controls (Keller 2014)
5. Correct inference
   - Robust / cluster / permutations / check for heteroschedasticity (Domingue 2022)
   - Multiple hypothesis testing

EMPIRICAL APPLICATION

E = School Starting Age (aug-sept cutoff)

G = Polygenic Index of Educational Attainment

Y = Test Scores

SCHOOL ENTRY POLICY

Fit continuous age into discrete grades → arbitrary cutoff for starting school

- UK: September 1st
    - Born Aug 31st: start 4yrs + 1 day
    - Born Sept 1st: start at 5yrs
- Consequence:
    - Age at tests
    - Developmental time spent at home
    - Relative age of students
- Policy: public edu reduces inequality?
- Family: “precocious” kid should start early?

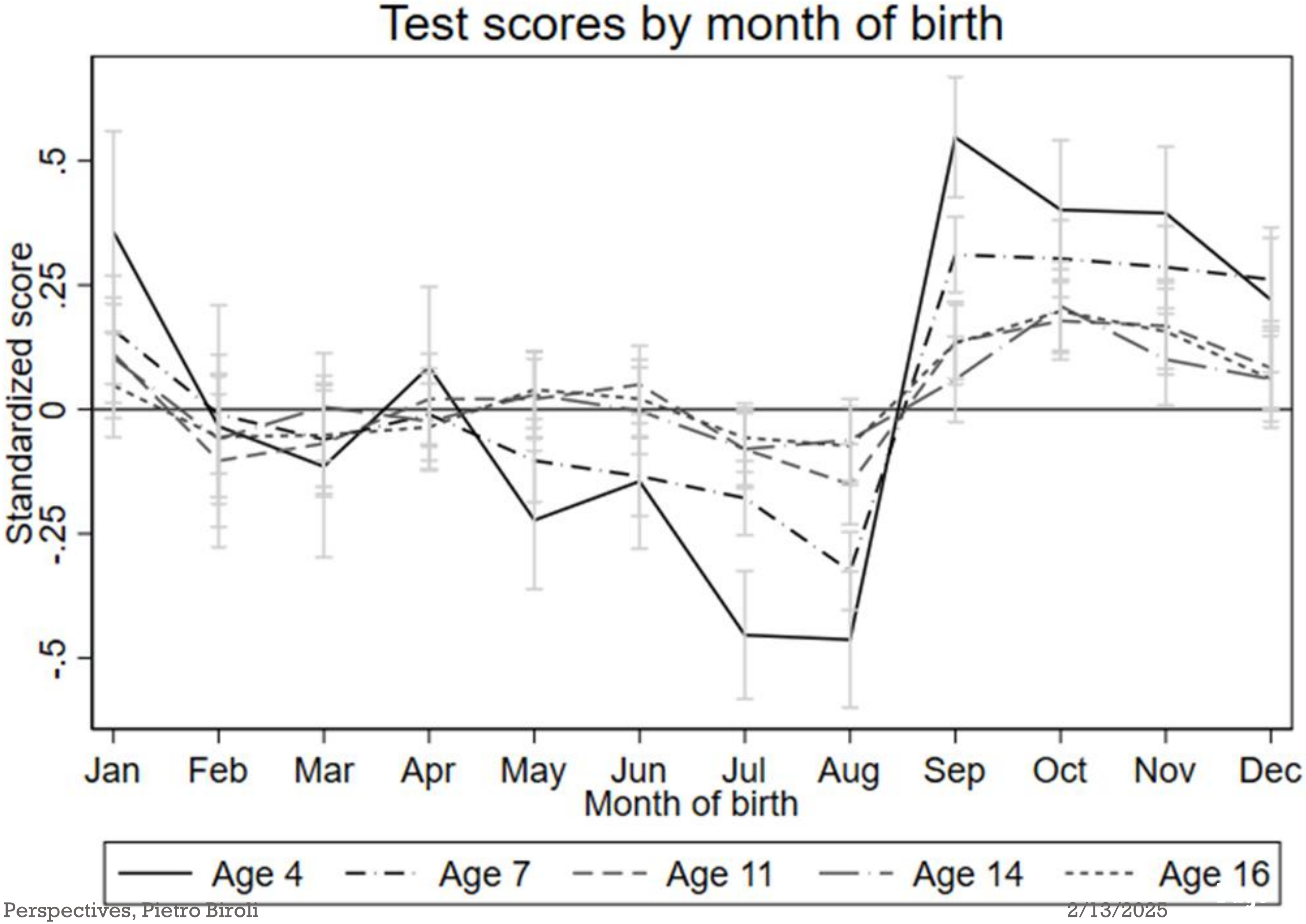
DATA

- Avon Longitudinal Study of Parents and Children (ALSPAC)
- Cohort study of preg women in 1991-1992 near Bristol
  - 14,541 pregnancies; 14,676 fetuses; 13,988 alive at age 1
  - About 3,500 have data on mom, dad, child genotype and tests
- Strict school starting age
- EA4-PGI
  - UKB-23&me sumstats; Ldpred
- 5 standardized exams (admin data)
  - Entry assessment (age 4), just before starting elementary school
  - Key Stage 1-2-3-4/GCSE (age 7-11-14-16), taken in class

TREATMENT EFFECT

RDD by month of birth

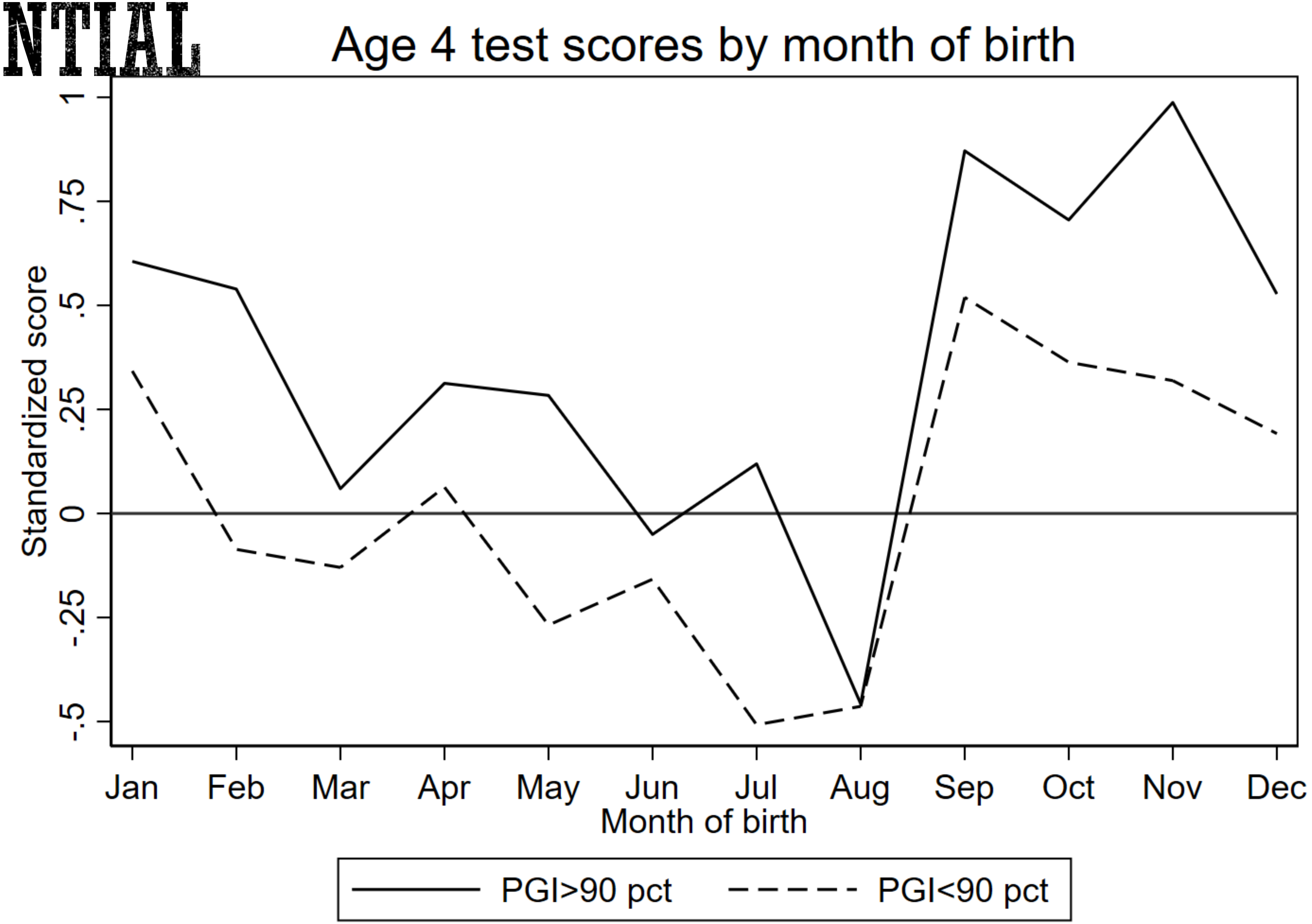
Treatment effect fading with age



GXE DIFFERENTIAL JUMP

RDD by high and
low
PGI

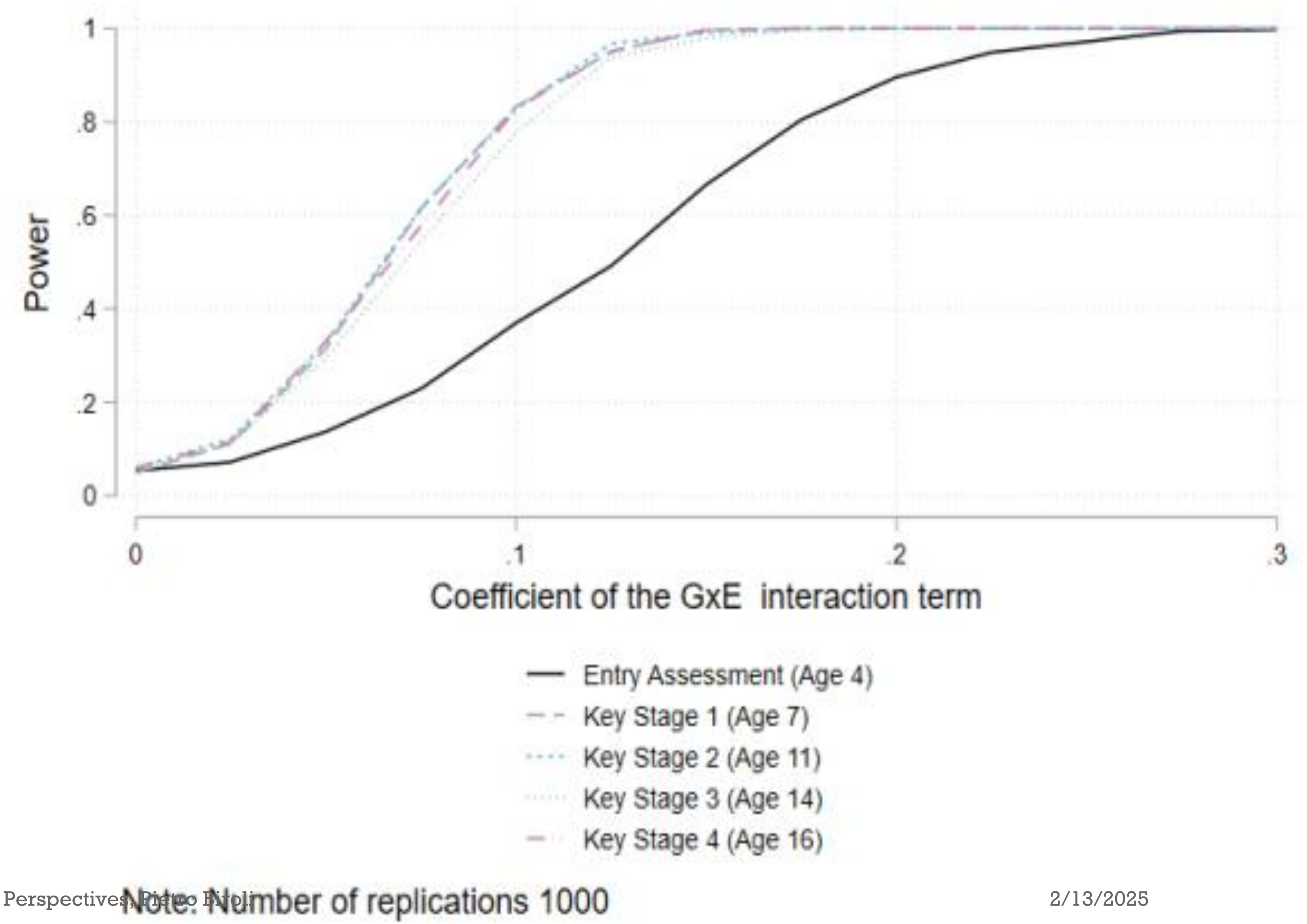
Age 4 test scores



1. POWER CALCULATIONS

Well powered (> 80%) to estimate an interaction coefficient:

- > 0.1 for Key Stage
- > 0.175 for the entry assessment



2. RGE

- PGI not associated to treatment

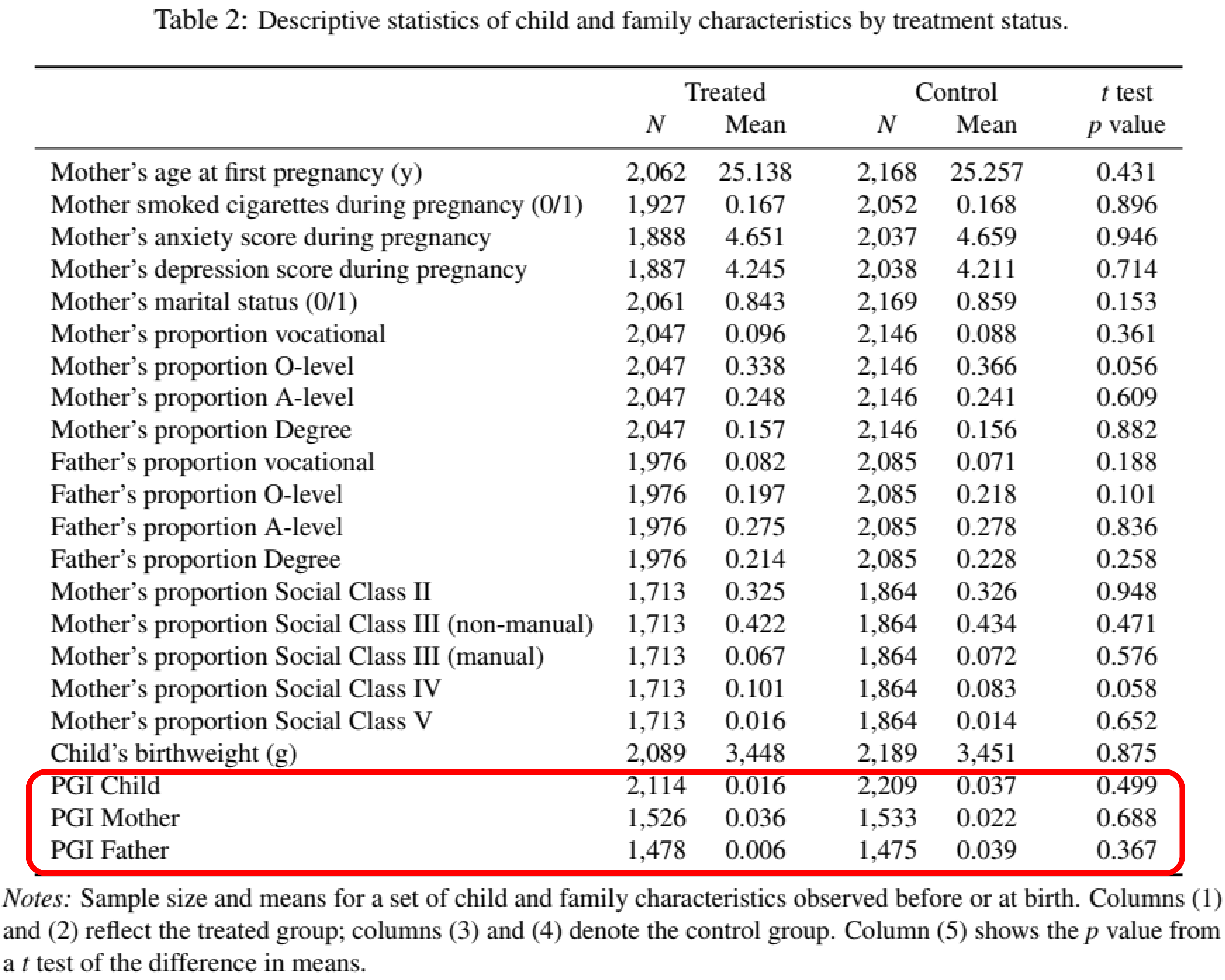


Table 2: Descriptive statistics of child and family characteristics by treatment status.

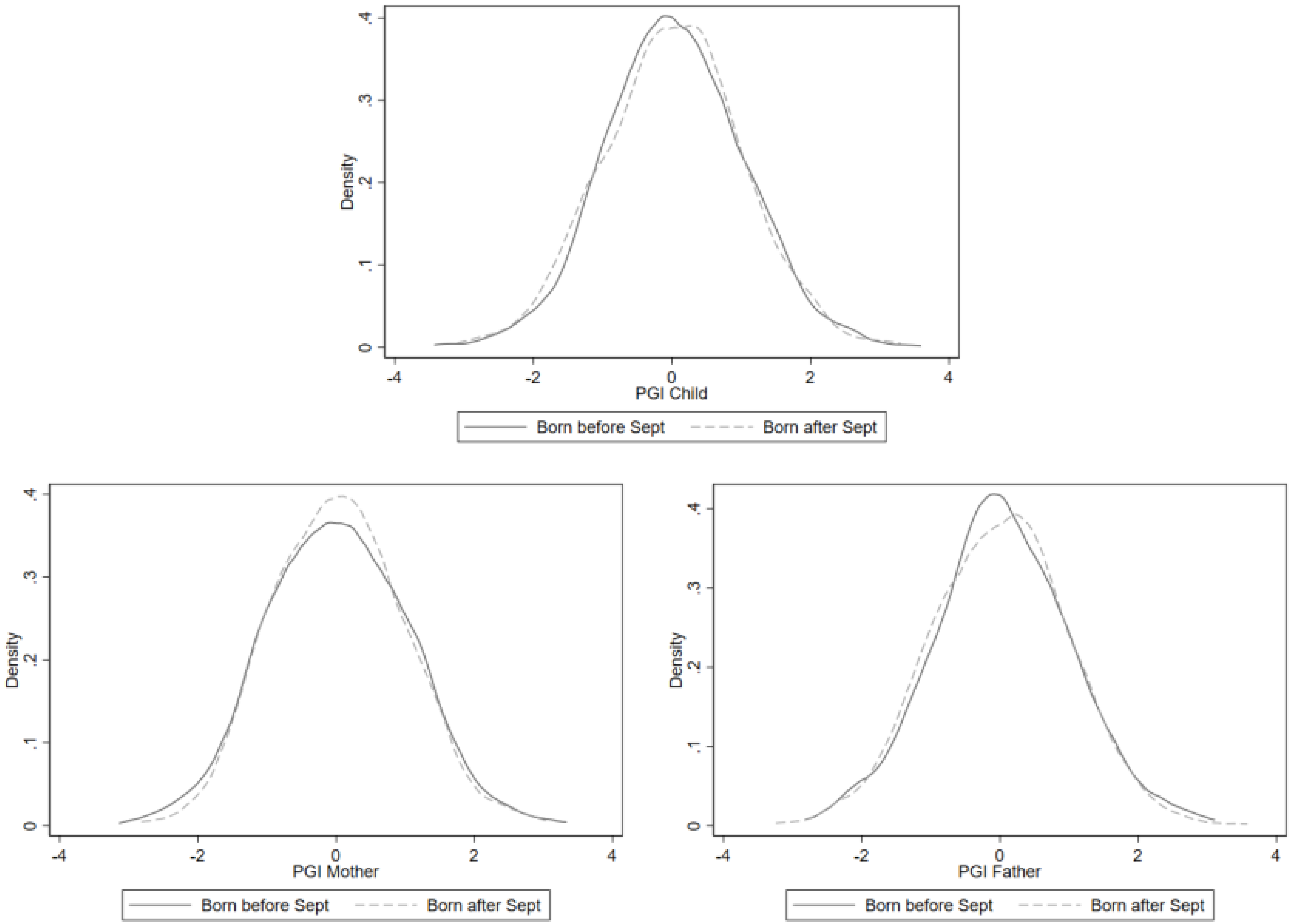
	Treated		Control		*t* test
	N	Mean	*N*	Mean	*p* value
Mother's age at first pregnancy (y)	2,062	25.138	2,168	25.257	0.431
Mother smoked cigarettes during pregnancy (0/1)	1,927	0.167	2,052	0.168	0.896
Mother's anxiety score during pregnancy	1,888	4.651	2,037	4.659	0.946
Mother's depression score during pregnancy	1,887	4.245	2,038	4.211	0.714
Mother's marital status (0/1)	2,061	0.843	2,169	0.859	0.153
Mother's proportion vocational	2,047	0.096	2,146	0.088	0.361
Mother's proportion O-level	2,047	0.338	2,146	0.366	0.056
Mother's proportion A-level	2,047	0.248	2,146	0.241	0.609
Mother's proportion Degree	2,047	0.157	2,146	0.156	0.882
Father's proportion vocational	1,976	0.082	2,085	0.071	0.188
Father's proportion O-level	1,976	0.197	2,085	0.218	0.101
Father's proportion A-level	1,976	0.275	2,085	0.278	0.836
Father's proportion Degree	1,976	0.214	2,085	0.228	0.258
Mother's proportion Social Class II	1,713	0.325	1,864	0.326	0.948
Mother's proportion Social Class III (non-manual)	1,713	0.422	1,864	0.434	0.471
Mother's proportion Social Class III (manual)	1,713	0.067	1,864	0.072	0.576
Mother's proportion Social Class IV	1,713	0.101	1,864	0.083	0.058
Mother's proportion Social Class V	1,713	0.016	1,864	0.014	0.652
Child's birthweight (g)	2,089	3,448	2,189	3,451	0.875
PGI Child	2,114	0.016	2,209	0.037	0.499
PGI Mother	1,526	0.036	1,533	0.022	0.688
PGI Father	1,478	0.006	1,475	0.039	0.367

Notes: Sample size and means for a set of child and family characteristics observed before or at birth. Columns (1) and (2) reflect the treated group; columns (3) and (4) denote the control group. Column (5) shows the *p* value from a *t* test of the difference in means.

2. RGE

- PGI not associated to treatment

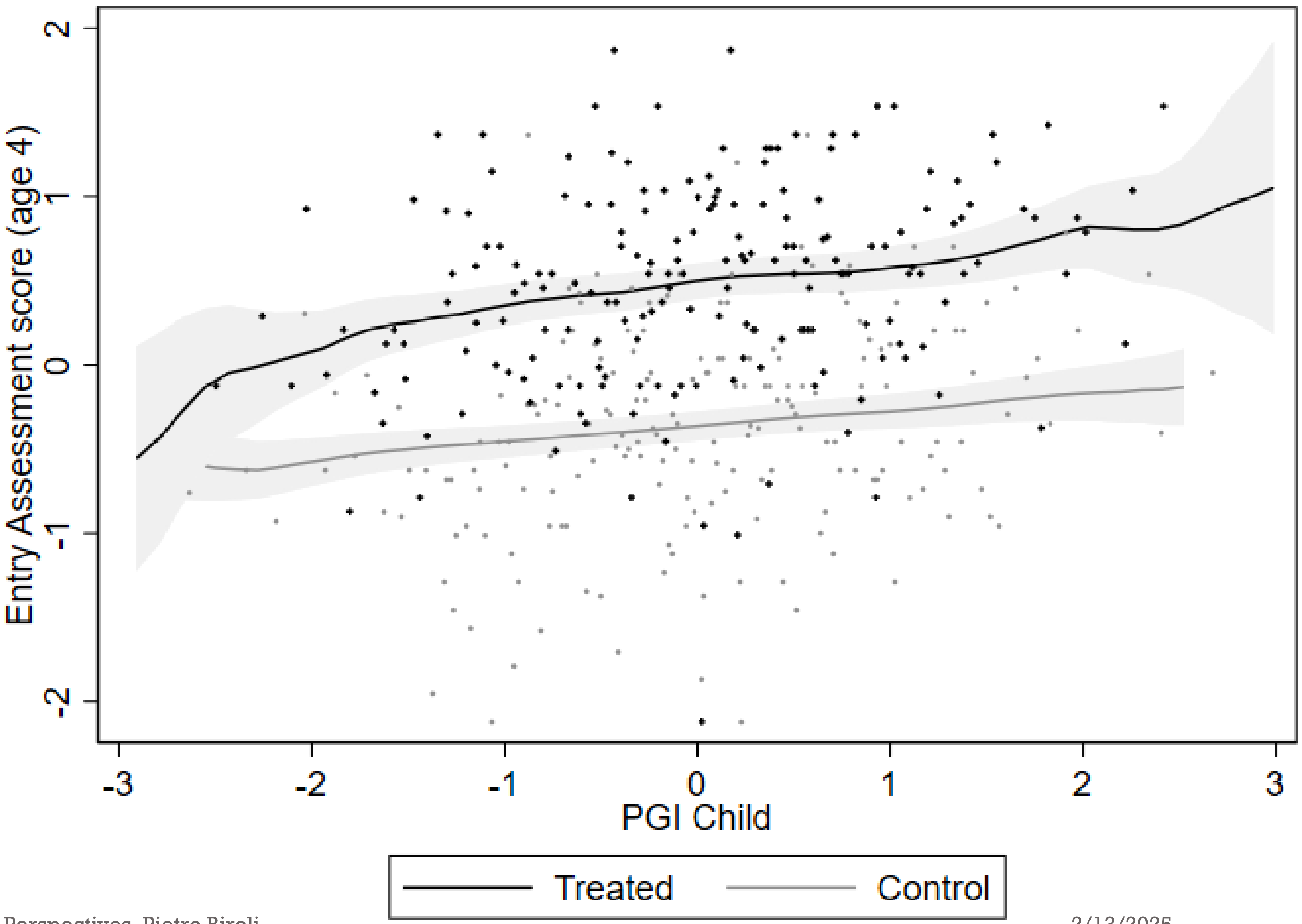
Figure G.1: Densities of children's, mothers' and fathers' PGI by treatment status.



3. FUNCTIONAL FORM

Evidence of

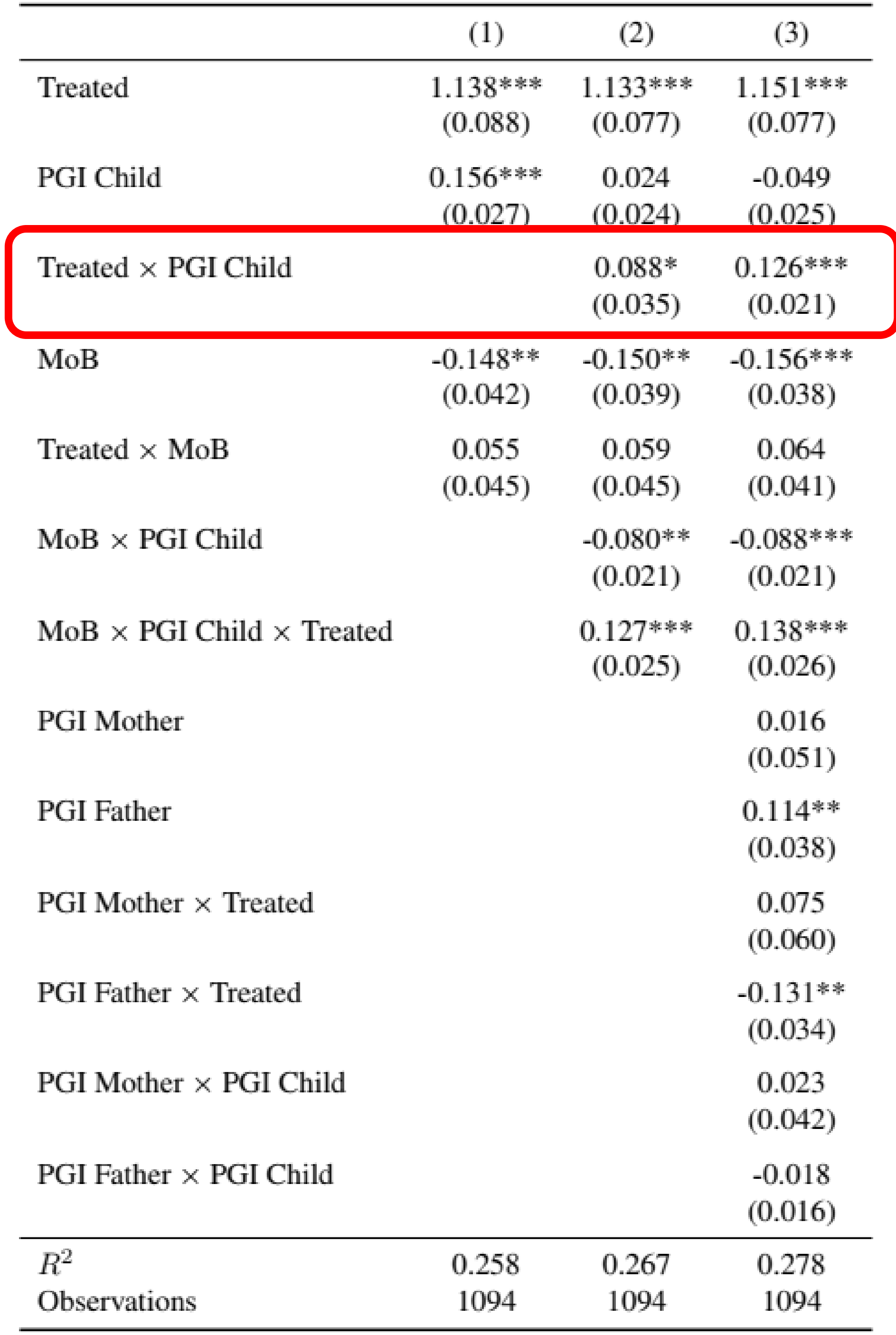
- Linearity
- Positive GxE



4. GXE AT AGE 4

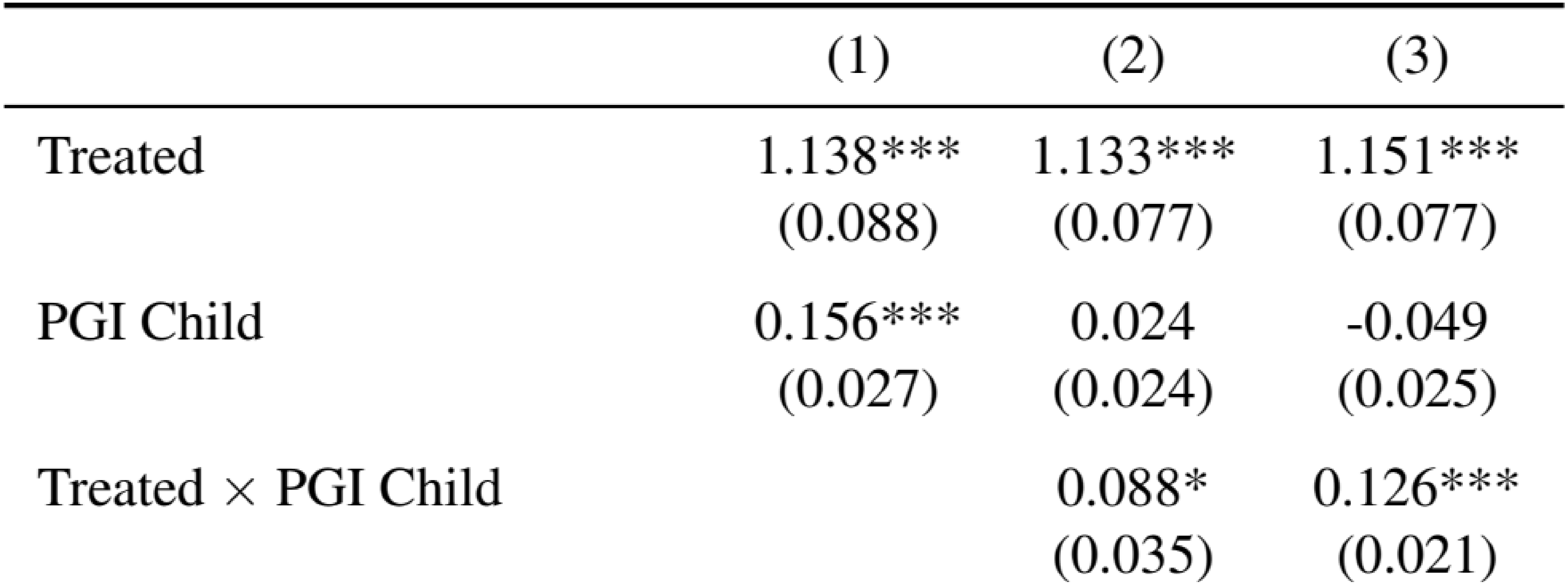
Entry assessment (age 4)

- Old-for-grade and high PGI perform better
- Positive interaction term:
  - high PGI benefit the most from spending 1 more year before starting school
  - Complementarity between parental investment and PGI
  - Robust to inclusion of parental PGI



	(1)	(2)	(3)
Treated	1.138*** (0.088)	1.133*** (0.077)	1.151*** (0.077)
PGI Child	0.156*** (0.027)	0.024 (0.024)	-0.049 (0.025)
Treated × PGI Child		0.088* (0.035)	0.126*** (0.021)
MoB	-0.148** (0.042)	-0.150** (0.039)	-0.156*** (0.038)
Treated × MoB	0.055 (0.045)	0.059 (0.045)	0.064 (0.041)
MoB × PGI Child		-0.080** (0.021)	-0.088*** (0.021)
MoB × PGI Child × Treated		0.127*** (0.025)	0.138*** (0.026)
PGI Mother			0.016 (0.051)
PGI Father			0.114** (0.038)
PGI Mother × Treated			0.075 (0.060)
PGI Father × Treated			-0.131** (0.034)
PGI Mother × PGI Child			0.023 (0.042)
PGI Father × PGI Child			-0.018 (0.016)
R^2	0.258	0.267	0.278
Observations	1094	1094	1094

GXE AT AGE 4



	(1)	(2)	(3)
Treated	1.138***	1.133***	1.151***
	(0.088)	(0.077)	(0.077)
PGI Child	0.156***	0.024	-0.049
	(0.027)	(0.024)	(0.025)
Treated × PGI Child		0.088*	0.126***
		(0.035)	(0.021)

CONCLUSION

GENE-ENVIRONMENT INTERPLAY

- Theoretically everywhere
- Empirically very hard to estimate

THANK YOU

INSTRUMENTAL VARIABLES ESTIMATION

THE TRADITIONAL VIEW: CONSTANT EFFECT MODEL

Suppose you have the following model

$$Y_i = \delta D_i + X_i \beta + \varepsilon_i$$

where Y_i is the outcome, D_i is the variable of interest, X_i is a vector of covariates including a constant, β is a vector of nuisance parameters, and ε_i is a regression error with mean zero and $cov(X_i, \varepsilon_i) = 0$.

Constant effect HP: β, δ are constants

- If $cov(D_i, \varepsilon_i) \neq 0$, then D_i is endogenous and OLS is inconsistent.
- Consider an example where the outcome is earnings, and D_i is number of children a woman has. Other covariates in X_i include education, experience, and husband's earnings. Why might number of children be endogenous?

Think of the variable D_i as being composed of two parts

$$D = b\varepsilon + c$$

where $cov(c, \varepsilon) \equiv 0$. Using this decomposition, we can write

$$Y = \delta c + X\beta + (1 - \delta b)\varepsilon$$

If we observed the different components of D_i, you could regress the outcome on c, which is the exogenous part of D_i, and get a consistent estimate of δ.

In reality, we do not observe the components of D_i, so the best thing we can do is to use an IV.

The idea behind IV is to find a variable which is correlated with c, the exogenous part of D_i, and is uncorrelated with ε

Z is an instrument for D when the following two conditions are met:

- Exclusion restriction: $Cov(Y, Z|X, D) = 0$.
  This implies that Z is exogenous, or $E(\varepsilon|Z) = 0$
- Instrument condition: $Cov(Z, D) \neq 0$. It must be correlated with D

If these hold, then we can use two-stage least squares (2SLS) and recover an unbiased estimate of δ

TWO STAGE LEAST SQUARES (2SLS)

In the first stage, we regress the endogenous variable D on all the exogenous variables including the instrument

$$D_i = a_0 + a_1 Z_i + a_2 X_i + u$$

We recover the predicted values, $\hat{D}$ of D from the first stage

In the second stage, we regress the outcome variable on the predicted $\hat{D}$ and the other Xs

$$y_i = a + \delta \hat{D}_i + X_i \beta + \varepsilon_i$$

$\hat{\delta}$ in the second stage is our IV estimate of the treatment impact

REDUCED FORM

Let's unpack the exclusion restriction: $Cov(Y, Z | X, D) = 0$.

- This means that Z can influence Y **only through** X
- What if there's another "path" W through which Z influences Y? Exclusion restriction does not hold any more!

- If it is still true that Z is exogenous [i.e. $E(\varepsilon | Z) = 0$] then you can run an OLS regression of Y and Z
- This is sometimes called the "reduced form"
  - Drawback: I cannot make inference of effect of X on Y, but only Z on Y